Computational Text Analysis

Friday 13:45 17:15 bi-weekly 2022 SOWI-Zoomroom 6

Marius Sältzer

marius.saeltzer@gesis.org
Office Hours: TBD

Syllabus will be adapted over the course of the semester.

Course Description:
Quantitative Text Analysis and Natural Language Processing are rapidly becoming staple tools in social science. The field is advancing rapidly, as new methods from computer science are spilling over and find application in traditional questions of political science, communication and sociology. This course gives an in-depth introduction to the core principles of text analysis, with a focus on the application of state of the art techniques in machine learning to individual research projects. Participants should be proficient in R, Python will be taught along the way. Coursework (2 coding tasks and 1 final paper) can be handed in relating to data relevant to individual research questions.

Formalia:

- This is a methods course and I want you to get access to the tools to collect your own data.
- Your Grade depends two things:
  - A *Term Paper* based on your research that connects to text analysis. I will review it journal style but grade it according to the fit to text analysis.
  - Homework: In the beginning of the course, I want you to choose a textual dataset from your own research area, that might interest you or will be part of your future research. The coding homework is mostly adapting the code to your own data and presenting relevant results.

I won't force you, but would like to encourage you to do

- *Presentation*: You will present your paper idea and get feedback by me...
- *Peer Review*: ...and you! You will add a review to your peers and discuss your papers.

- Software: R and Python
  - Basics: R
  - Advanced (multilingual, transformers): Python
  - Script-based - you will get scripts ready to run, but have to adapt them for homework.
- Helpful Textbooks:
  - NLP
    * Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* URL `https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf`
    * https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf
    * J.j Allaire. *Deep Learning with R.* Manning Publications Co. LLC, New York, 2018. ISBN 9781638351634
  - Content Analysis: Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Sage Publ, Thousand Oaks, Calif., 2. ed., [nachdr.] edition, 2009. ISBN 0-7619-1545-1

Session 1 (25.02.2022) : Introduction:

Week	Content
Introduction	1. Motivation & Discipline 2. What Text Analysis can do 3. Content Analysis (a) Human Understanding of Text (b) Context (c) Concepts (d) Encoding Text 4. Natural Language Processing (a) Linguistics (b) Grammar (c) Semantics
Introduction to Quanteda	1. Regular Expressions 2. Example: A simple Dictionary 3. Corpora 4. Tokenization
Further Reading	Literature: • Krippendorff 2009: **(Chapter 1 & 4)** • Additional: – Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013. ISSN 1047-1987. doi: 10.1093/pan/mps028 – Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as Data. *Journal of Economic Literature*, 57(3):535–574, 2019. ISSN 0022-0515. doi: 10.1257/jel.20181020 Homework: Find your dataset!

Session 2 (11.03.2022): Text as Data:

Week	Content
Bag of Words	1. Document-Feature-Matrix 2. The Curses of the Natural Language 3. Preprocessing with Features (a) Stemming (b) Lemmatization (c) N-Grams
Context is for Kings	Word Embeddings 1. The logic of Word Embeddings 2. Word-to-Vec 3. Doc-to-Vec
Preparation	Literature: • https://tutorials.quanteda.io/basic-operations • Jurafsky & Martin (2020) : Chapter 2, 3 & 6 • Additional: – Austin C. Kozlowski, Matt Taddy, and James A. Evans. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949, 2019. ISSN 0003-1224. doi: 10.1177/0003122419877135 – Jack Blumenau. The Effects of Female Leadership on Women's Voice in Political Debate. *British Journal of Political Science*, 51(2):750–771, 2021. ISSN 0007-1234. doi: 10.1017/S0007123419000334 Data: • TBD Homework: • Homeworkscript "Text to Data" • Session2.R

Session 3 (25.03.2022): Unsupervised Learning:

Week	Content
Distributional Semantics	1. Zipf's Law 2. Semantic Markers for Concepts 3. tf-idf
Topic Models	1. LDA Topic Models 2. Interpreting Topic Models
LSA & Friends	1. Latent Semantic Analysis 2. Correspondence Analysis 3. Wordfish
Preparation	Literature: • David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55 (4):77–84, 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826 • Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008. ISSN 00925853. doi: 10.1111/j.1540-5907.2008.00338.x • Additional: – Alexander Baturo, Niheer Dasandi, and Slava Mikhaylov. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research and Politics*, (2), 2017. doi: 10.1177/20531680177128 – Pablo Barberá, Andrue Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901, 2019. ISSN 0003-0554. doi: 10.1017/S0003055419000352 – Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. ISSN 00925853. doi: 10.1111/ajps.12103 Data Suggestions: • Manifestos • United Nations Speeches Homework: Script 3

Session 4 (08.04.2022): Supervised Learning:

Week	Content
Supervised Learning	1. Concept 2. Principles 3. Validation
Bag-of-Words Classifier	1. Logistic Regression 2. Naive Bayes 3. Support-Vector Machines
Deep Learning	1. Concepts 2. Introduction to Keras
Preparation	Literature: • Stefan Müller. The Temporal Focus of Campaign Communication. *The Journal of Politics*, 84(1):585–590, 2022. ISSN 0022-3816. doi: 10.1086/715165 • Vladislav Petkevic and Alessandro Nai. Political Attacks in 280 Characters or Less: A New Tool for the Automated Classification of Campaign Negativity on Social Media. *American Politics Research*, page 1532673X2110556, 2021. ISSN 1532-673X. doi: 10.1177/1532673X211055676 • Jurasfsky & Martin (2020) : Chapter 4 & 6 Data: • https://manifestoproject.wzb.eu/information/documents/manifestoR Homework: Script 4

Session 5 (22.04.2022): Dealing with Real Data:

Week	Content
16.11.2022	1. Text in the Wild (a) Length: Survey Responses (b) Jargon: Legal Text (c) Noise: Twitter (d) Institutional Constraints: Parliamentary Speeches (e) Problematic Labels: Election Manifestos
Taming Text	1. Finding Labels (a) Dictionaries (b) Pretrained Models 2. Labeling Data yourself
Preparation	Literature: • Sven-Oliver Proksch and Jonathan B. Slapin. How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany. *German Politics*, 18 (3):323–344, 2009. doi: 10.1080/09644000903055799 • Slava Mikhaylov, Michael Laver, and Kenneth R. Benoit. Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis*, 20(1):78–91, 2012. ISSN 1047-1987. doi: 10.1093/pan/mpr047 • Kenneth Benoit, Michael Laver, and Slava Mikhaylov. Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science*, 53(2):495–513, 2009. ISSN 00925853 • Additional Reading for Manual Coding – – Krippendorff 2009: **Chapter 5, 6** Data: • Christian Rauh, Pieter de Wilde, and Jan Schwalbach. *The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states.* Center for Open Science, 2017. doi: 10.31235/osf.io/c5gdm

Session 6 (06.05.2022): You Presentations:

Week	Content
Presentations	I will assign 2 reviewers per paper, so you will get 3 reviews! I will try to cluster people that share a field to get some relevant perspectives. 1. Student 2. Student 3. Student 4. Student 5. Student 6. Student 7. Student 8. Student 9. Student 10. Student 11. Student 12. Student 13. Student 14. Student 15. Student

Session 7 (20.05.2022: Advanced Solutions:

Week	Content
TBD	A selection based on your needs: 1. Mulitlingual Embeddings 2. Transfer Learning 3. Named entity recognition 4. Latent Semantic Scaling 5. POS-tagging 6. Transformer Models
Preparation	Literature: • Data: • •

References

J.j Allaire. *Deep Learning with R*. Manning Publications Co. LLC, New York, 2018. ISBN 9781638351634.

Pablo Barberá, Andrue Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901, 2019. ISSN 0003-0554. doi: 10.1017/S0003055419000352.

Alexander Baturo, Niheer Dasandi, and Slava Mikhaylov. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research and Politics*, (2), 2017. doi: 10.1177/20531680177128.

Kenneth Benoit, Michael Laver, and Slava Mikhaylov. Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science*, 53(2): 495–513, 2009. ISSN 00925853.

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826.

Jack Blumenau. The Effects of Female Leadership on Women's Voice in Political Debate. *British Journal of Political Science*, 51(2):750–771, 2021. ISSN 0007-1234. doi: 10.1017/S0007123419000334.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as Data. *Journal of Economic Literature*, 57(3):535–574, 2019. ISSN 0022-0515. doi: 10.1257/jel.20181020.

Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013. ISSN 1047-1987. doi: 10.1093/pan/mps028.

Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* URL `https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf`.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949, 2019. ISSN 0003-1224. doi: 10.1177/0003122419877135.

Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Sage Publ, Thousand Oaks, Calif., 2. ed., [nachdr.] edition, 2009. ISBN 0-7619-1545-1.

Slava Mikhaylov, Michael Laver, and Kenneth R. Benoit. Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis*, 20(1):78–91, 2012. ISSN 1047-1987. doi: 10.1093/pan/mpr047.

Stefan Müller. The Temporal Focus of Campaign Communication. *The Journal of Politics*, 84(1): 585–590, 2022. ISSN 0022-3816. doi: 10.1086/715165.

Vladislav Petkevic and Alessandro Nai. Political Attacks in 280 Characters or Less: A New Tool for the Automated Classification of Campaign Negativity on Social Media. *American Politics Research*, page 1532673X2110556, 2021. ISSN 1532-673X. doi: 10.1177/1532673X211055676.

Sven-Oliver Proksch and Jonathan B. Slapin. How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany. *German Politics*, 18(3):323–344, 2009. doi: 10.1080/09644000903055799.

Christian Rauh, Pieter de Wilde, and Jan Schwalbach. *The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states.* Center for Open Science, 2017. doi: 10.31235/osf.io/c5gdm.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. ISSN 00925853. doi: 10.1111/ajps.12103.

Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008. ISSN 00925853. doi: 10.1111/j.1540-5907.2008.00338.x.